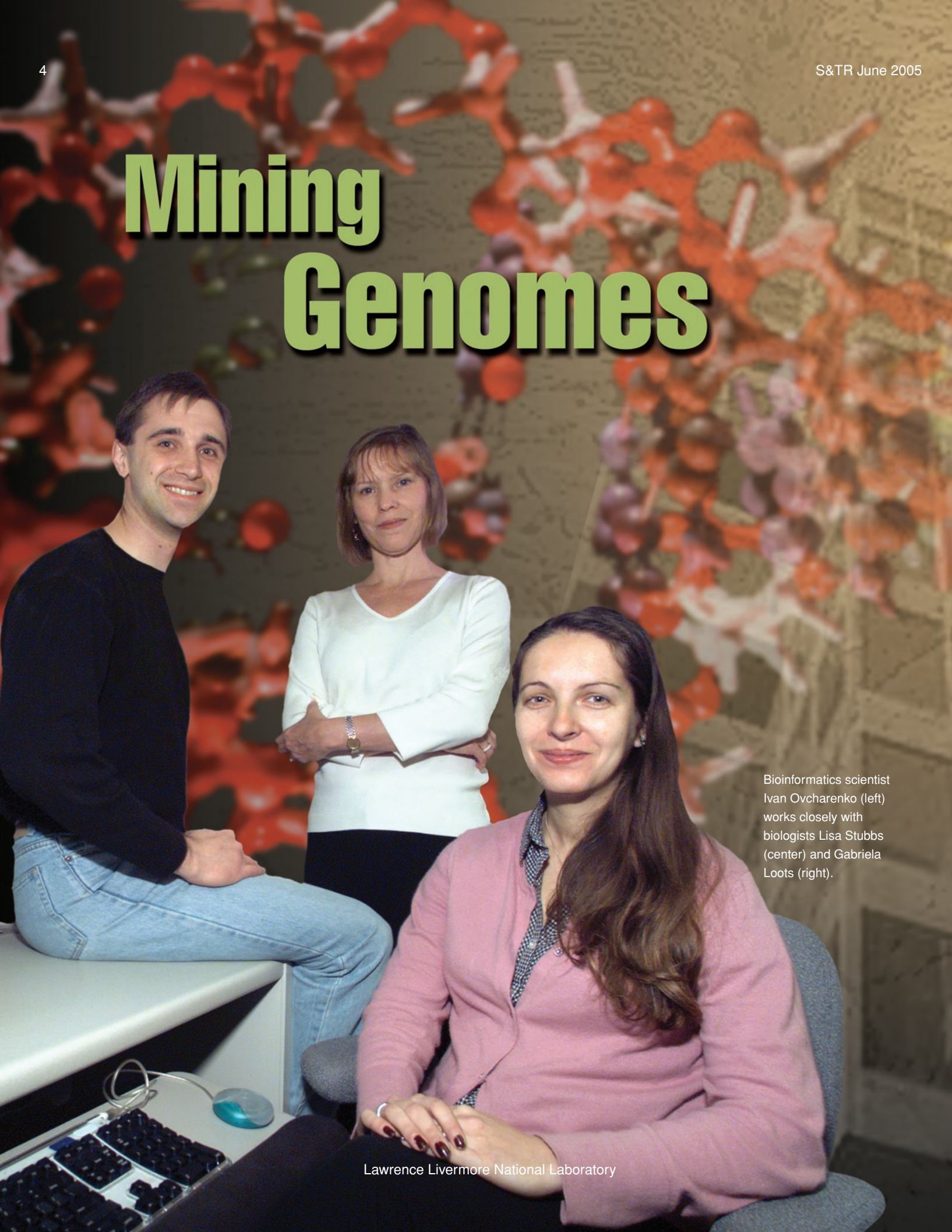
# Mining Genomes



Bioinformatics scientist Ivan Ovcharenko (left) works closely with biologists Lisa Stubbs (center) and Gabriela Loots (right).

***Livermore computational tools are helping researchers worldwide understand how genes are regulated.***

THE completion of the Human Genome Project in 2003 marked the most ambitious research effort in the history of life sciences: the sequencing of human DNA. However, the project, which included the participation of Lawrence Livermore biologists and computational scientists, was only the first step in understanding life at the molecular level. "The Human Genome Project gave us the sequence of the human DNA but not the manual that explains what it means," says Ivan Ovcharenko, a bioinformatics scientist in Livermore's Computation Directorate.

A significant challenge remains in identifying regulatory elements (REs), which are sequences of DNA that interact with specific proteins to serve as "on–off" switches for genes. REs function in remarkably similar ways in all species, from microbes to humans. However, finding them is particularly challenging in large and complex genomes such as those of mammals. Human REs can be located directly next to the gene they control, or they can be 1-million DNA bases away, buried in a "gene desert." Gene deserts, of particular interest to many biologists, are the large and seemingly barren areas located between genes.

Gene deserts were once considered to have no biological use and were dismissed as "junk DNA." (See the box below.) However, researchers at Lawrence Livermore and Lawrence Berkeley national laboratories and the Department of Energy's Joint Genome Institute (JGI) have shown that gene deserts contain large numbers of REs. Medical and

## A Short Genome Primer

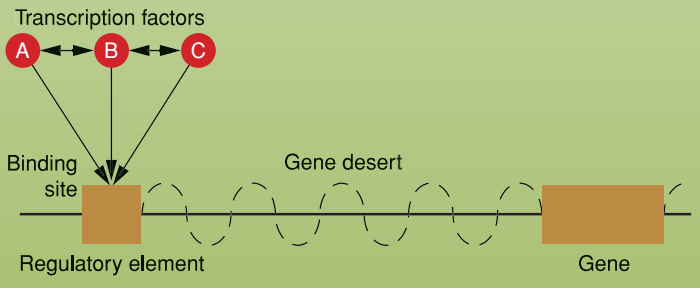
The genome is an organism's complete set of DNA. Genomes vary widely in size: the smallest known genome for a bacterium contains about 600,000 DNA base pairs, while human and mouse genomes have some 3 billion base pairs, equivalent to 3 gigabytes of information. Except for mature red blood cells, all human cells contain a complete genome. Each cell contains 23 pairs of chromosomes in their nuclei, and each chromosome contains two tightly coiled strands of DNA (deoxyribonucleic acid). A DNA strand is an extremely long polymer composed of sequences of nucleotides—the four chemical bases of adenine, cytosine, guanine, and thymine, commonly abbreviated as A, C, G, and T. These chemical bases are attached to a sugar–phosphate chain. The DNA sequence is the particular side-by-side arrangement of bases along the DNA strand (for example, ATTCCGGA).

Certain sequences of letters constitute genes, which are responsible for making proteins. Genes differ only in their sequence of DNA bases. Surprisingly, the 30,000 genes (each about 3,000 letters long) that have been identified through the Human Genome Project comprise only 2 percent of the human genome. About 45 percent of the genome is repetitive DNA that has accumulated over millions of years of evolution and seems to serve no function.

The remaining 53 percent of the genome is noncoding, nonrepetitive DNA, the function of which is still unclear. Part of the noncoding DNA is composed of so-called gene deserts, long stretches of sequences without any protein-coding genes. These gene deserts can stretch more than 5 million bases in length between genes. Some deserts are home to regulatory elements (REs), which are sequences of DNA about 200 letters long and are involved in turning genes "on" and "off." REs combine with transcription factors, which are small proteins, at specific binding sites. When this happens, the DNA strand unravels so that the RE–transcription factor complex abuts the gene it regulates, causing the gene to begin the process for making a certain protein.

A few institutions, including Lawrence Livermore and Lawrence Berkeley national laboratories, are leading the scientific efforts to study REs and determine their role in gene activity and disease. In 2003, a team of researchers from Lawrence Berkeley and the Department of Energy's Joint Genome Institute (JGI) compared DNA sequences from gene deserts in the human, mice, frog, and fish genomes and discovered sequences that regulate genes over surprisingly long distances. "Gene deserts may not be home to any genes, but they can host DNA sequences that act as long-distance switches to activate far away genes," says JGI Director Edward Rubin, who led this research. One of the coresearchers of this pioneering work was bioinformaticist Ivan Ovcharenko, who is now at Livermore.

"It appears that one RE can affect several genes, and one gene can have many REs," says Livermore biologist Gabriela Loots. "We're at the frontier of a new field." In time, scientists may also discover other important sequences in gene deserts.




Regulatory elements are found along vast stretches of gene deserts. They combine with transcription factors to regulate protein-coding genes.

biological researchers can learn important information about the origins and mechanisms of disease, including heart disease, cancer, and AIDS, from studying the functions of REs in these vast desert regions.

"Although REs exist as oases of function in deserts of nonfunction," says Ovcharenko, "it isn't clear how to reliably identify REs because they have no real signature." In an effort to help researchers locate REs in published genome sequences of many different species, Ovcharenko and colleagues have developed a suite of analytical and visualization computational tools. (See the box on p. 8.) The tools are part of Livermore's virtual Comparative Genomics Center, and researchers worldwide can access them online at www.dcode.org. The tool development team includes Ovcharenko, biologists Lisa Stubbs and Gabriela Loots of Livermore, Marcelo Nobrega of Lawrence Berkeley, and Ross Hardison and Webb Miller of Pennsylvania State University.

"Ivan is at the computational end, creating new methods and tools to tackle critical problems in genome biology," says Stubbs. "We provide the biological expertise to test their reliability and usefulness. After Ivan's programs make a set of predictions, we test those predictions in the laboratory to study RE function in live cells or in animals. (See *S&TR,* April 2005, pp. 20–22.) Ivan, in turn, takes our experimental results and refines his programs to enhance their performance."

"While the tools aren't 100 percent accurate, they are a significant advance that allows us to expedite our genomic research," says Loots. "Instead of going on a hunting expedition, we use the tools to zero in on stretches of DNA that offer the most potential."

## Genome Data Growing Fast

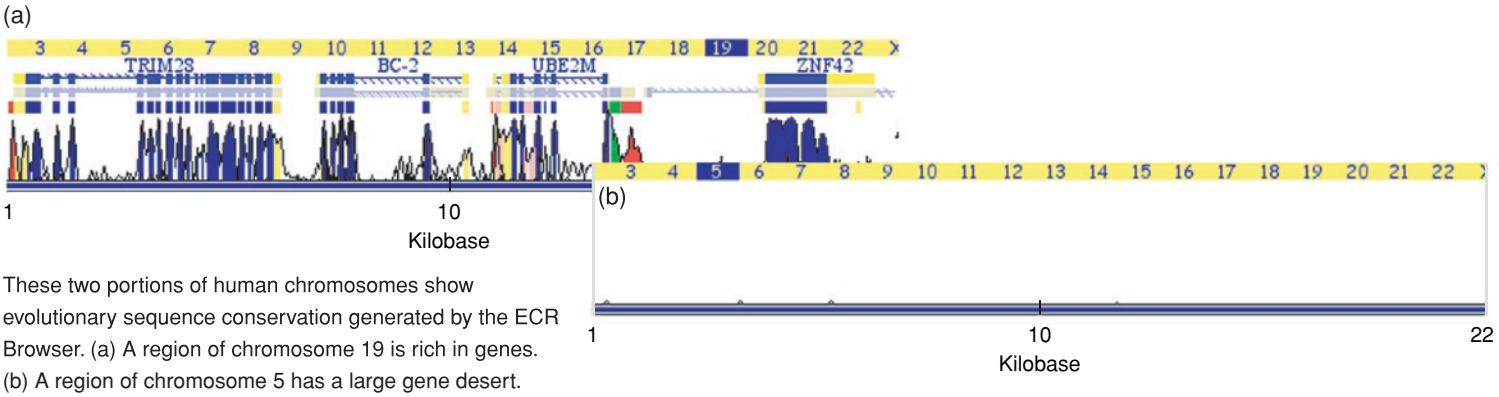
The Livermore suite of computational tools has become popular because of the rapidly growing body of sequenced genomes being generated in the public domain. "Biologists need fast and reliable methods and tools to efficiently analyze the massive amount of data emanating from public sequencing efforts," says Ovcharenko. For example, the sequencing of the chicken genome was completed last year, in an international effort involving Stubbs and her group at Livermore, and Susan Lucas and her group at JGI. The two groups sequenced portions of chromosome 11 and all of chromosome 28 in both domestic and wild chickens. The chicken genome sequence is important to biologists because of the animal's prominent role in agriculture, its major role as an experimental model for vertebrate development, and the strategic evolutionary position of birds between mammals and fish.

The computational tools are Web accessible and can be easily used by researchers who have minimal or no computer training. As a result, some 300 international users, with interests ranging from ecology to agriculture to clinical medicine, use the Livermore tools every day.

The team also strives to make it easier for users to visualize their results. "It's difficult to comprehend the meaning of millions of ordered letters, so we've created graphical interfaces to make visual sense of all the information," says Ovcharenko. For example, some programs generate phylogenetic trees that clearly depict the evolutionary relationship between species over millions of years and when one species likely diverged from another. The Livermore programs offer more than static electronic files of the results; they also provide dynamic and interactive data analysis.

"Researchers can walk into my laboratory one day and be using these tools productively the next," says Stubbs. "The Livermore tools have put powerful predictive methods into the hands of the people with the right scientific questions, and this capability has greatly accelerated the pace of discovery."

The tools align and analyze stretches of DNA or even entire genomes of species to find sequences of nucleotides—the four chemical bases of adenine (A), cytosine (C), guanine (G), and thymine (T)—that match exactly or have strong similarities.




These two portions of human chromosomes show evolutionary sequence conservation generated by the ECR Browser. (a) A region of chromosome 19 is rich in genes. (b) A region of chromosome 5 has a large gene desert.

These similar stretches are called evolutionary conserved regions (ECRs) because they represent areas of DNA that have mutated at much slower rates than the rest of the genome over tens and sometimes hundreds of millions of years. The strategy of comparing DNA from different species is called comparative genomics. Livermore researchers have used comparative genomics for several years to identify shared "core" functional elements that define fundamental properties, as well as genes and REs, of different species.

Comparisons between distantly related organisms, such as primates and fish, uncover the fundamental genomic building blocks—protein-coding genes and REs—that are shared by all vertebrates. Comparisons between more closely related species, such as mice and rats, can highlight genes and REs that are changing rapidly and define species-specific functions. This approach is particularly useful, for example, when studying human diseases that do not afflict other species. When mouse and human genomes are compared, about 40 percent of the two genomes are similar; when human and ape genomes are compared, more than 95 percent of the genomes are similar. In contrast, the sequences of humans and fish have diverged so significantly that only about 5 percent of those genomes are obviously related.

## Suite of Programs

The Livermore suite of DNA sequence-alignment programs includes zPicture, Mulan, and the ECR Browser: zPicture compares sequences of two species; Mulan compares sequences of multiple species; and ECR Browser compares complete genomes of many species.

The zPicture tool analyzes a chosen part of two species' genomes to identify ECRs. This tool is often used to compare nonvertebrate genomes, including those of microbes. Ovcharenko designed the genome alignment visualization so that the reference DNA sequence is linear along the horizontal axis and the percentage of similarity is plotted along the vertical axis. This arrangement provides an immediate visual annotation of candidate REs found in genomic sequences up to 1 million bases long.

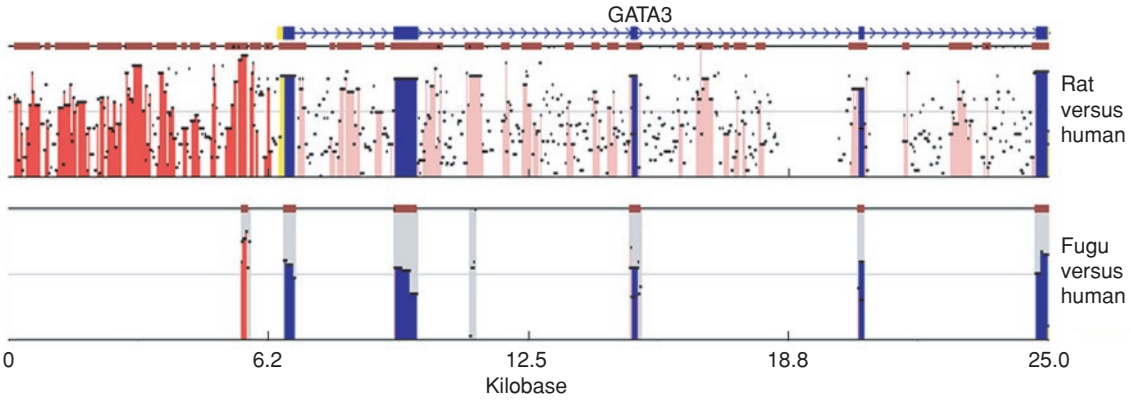
Mulan goes a step further than zPicture by aligning sequences from multiple species. Mulan determines the phylogenetic relationships among species and generates phylogenetic trees. It also provides detailed overviews of sequences, constructs DNA alignments in both graphic and textual formats, and presents users with several visual display options. The speed with which Mulan handles genomic sequences of millions of bases and the dynamic character of the user interface are unique among genomic tools.

The Livermore team illustrated the power of Mulan when the program analyzed the GATA3 gene and identified its potential REs. GATA3 is one of the key genes involved in the formation of bones, hair, and teeth. The team found that the gene was conserved among humans,

```
Human   GTGTATTgAACAaTaaGAGATAATAATCTAttaaCATTgTCaTcaCGTtGcGtTtTgCTC
        ||||||| ||||  | ||||||||||||||     |||| ||  |  ||| | | | | ||||
Mouse   GTGTATTtAACAcT--GAGATAATAATCTAaggcCATTtTCtTggCGT-GtGaTgTcCTC

Human   TGCCCtTcCaGacaTCTctACATGgAtGCCATaaGCTCtT-TCtTCTTATCTAGGTGTTg
        |||||| | |   ||| |||||  | |||||  |||| | ||  ||||||||||||||||
Mouse   TGCCCaTaCtG-tgTCTgcACATGtAaGCCATggGCTCcTgTCcTCTTATCTAGGTGTTt
```

Biologists use a suite of computational tools for aligning and analyzing stretches of DNA to compare sequences of nucleotides (the four chemical bases commonly abbreviated as A, C, G, and T). In this comparison of human and mouse genomes, the lower-case letters indicate mismatches; upper-case letters indicate those that match exactly or have strong similarities.




Analysis of the human, rat, and fugu (a Japanese pufferfish) GATA3 gene using the zPicture DNA sequence-alignment program shows much greater similarities in the DNA between rat and human than between fugu and human. Blue corresponds to the gene, and light red and dark red indicate noncoding DNA.

rodents, birds, amphibians, and fish, a group of species that span 450 million years of evolution.

The team also found five candidate REs that may regulate GATA3. One of them is present in all species, suggesting this RE plays a key role in GATA3. Three of the other REs are present in the human, rodent, and chicken genomes but are not detected in the frog and fish genomes. "One could speculate that the key involvement of the GATA3 gene in the growth of hair and feathers might be regulated by one of these three REs, and that their absence from the frog and fish genomes may be linked to the

## Marrying Computer Science and Biology

Finding regulatory elements on vast stretches of gene deserts would be impossible without sophisticated computer programs. The suite of programs developed by Ivan Ovcharenko and other Livermore researchers is indicative of the growing marriage between bioscience and computer science.

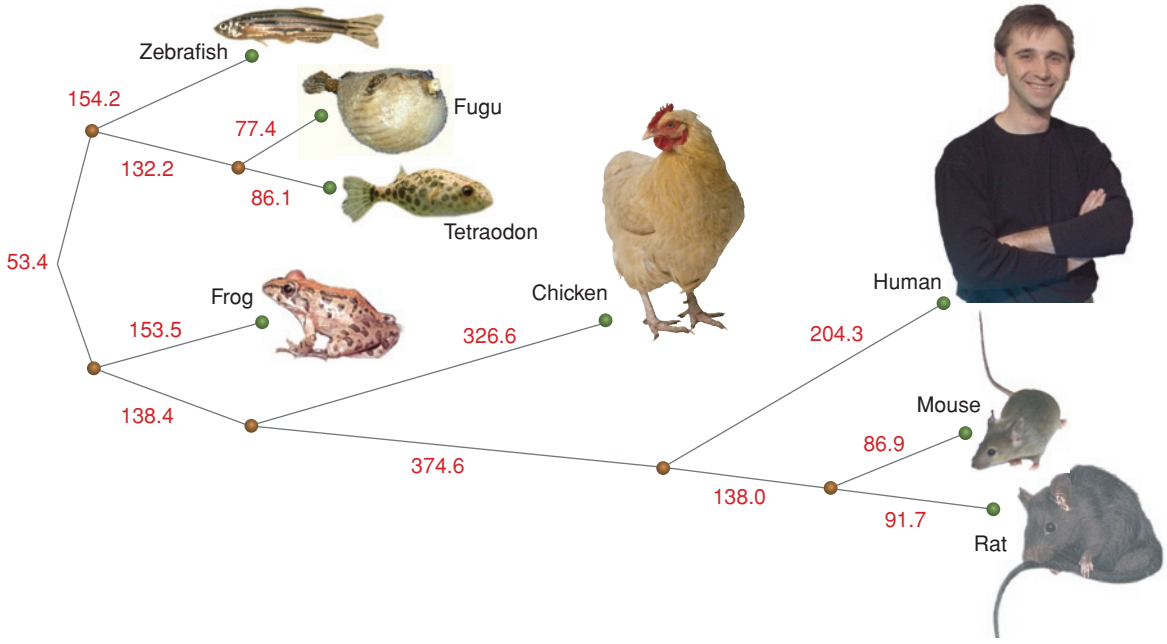
Peg Folta, division leader of Computer Applications Research for Energy, Environment, and Biological Computations, points out that computer scientists partnering with the Biosciences Directorate increasingly have a strong understanding of biology. "The blend of disciplines is becoming more commonplace, and we think it represents the future of biological research," she says.

Folta notes that computer science has been applied to the field of physics for many years. As a result, the field of computational physics is mature. However, this is not the case in computational biology. Until recently, the merging of the biology and computation fields was largely happenstance. A few computational scientists learned biology on their own to make important contributions, and a few biologists became computer savvy and learned to develop software and use computers to aid their research.

Applying computational science to biological challenges has matured in the past decade, especially with the completion of the Human Genome Project in 2003 and the resulting onslaught of information. The project would have been impossible to complete without powerful computers and specialized software. An increasing number of colleges offer both undergraduate and graduate degrees in computational biology or bioinformatics. Computational biologists analyze biological activity of living organisms, including protein–protein interactions, protein structures, and signaling networks using computational methods. Bioinformaticists decode the content encrypted into genomes, developing algorithms and tools to access, mine, and display genome information.

Folta cites a "closed loop" that currently characterizes much biological research: Biologists conduct laboratory experiments, and computer scientists analyze the resulting data by accessing online databases, developing specialized software, and generating models and simulations that give insight to the biologist to design new laboratory experiments. With each loop, the computational methods and software are refined and biological discovery increases.

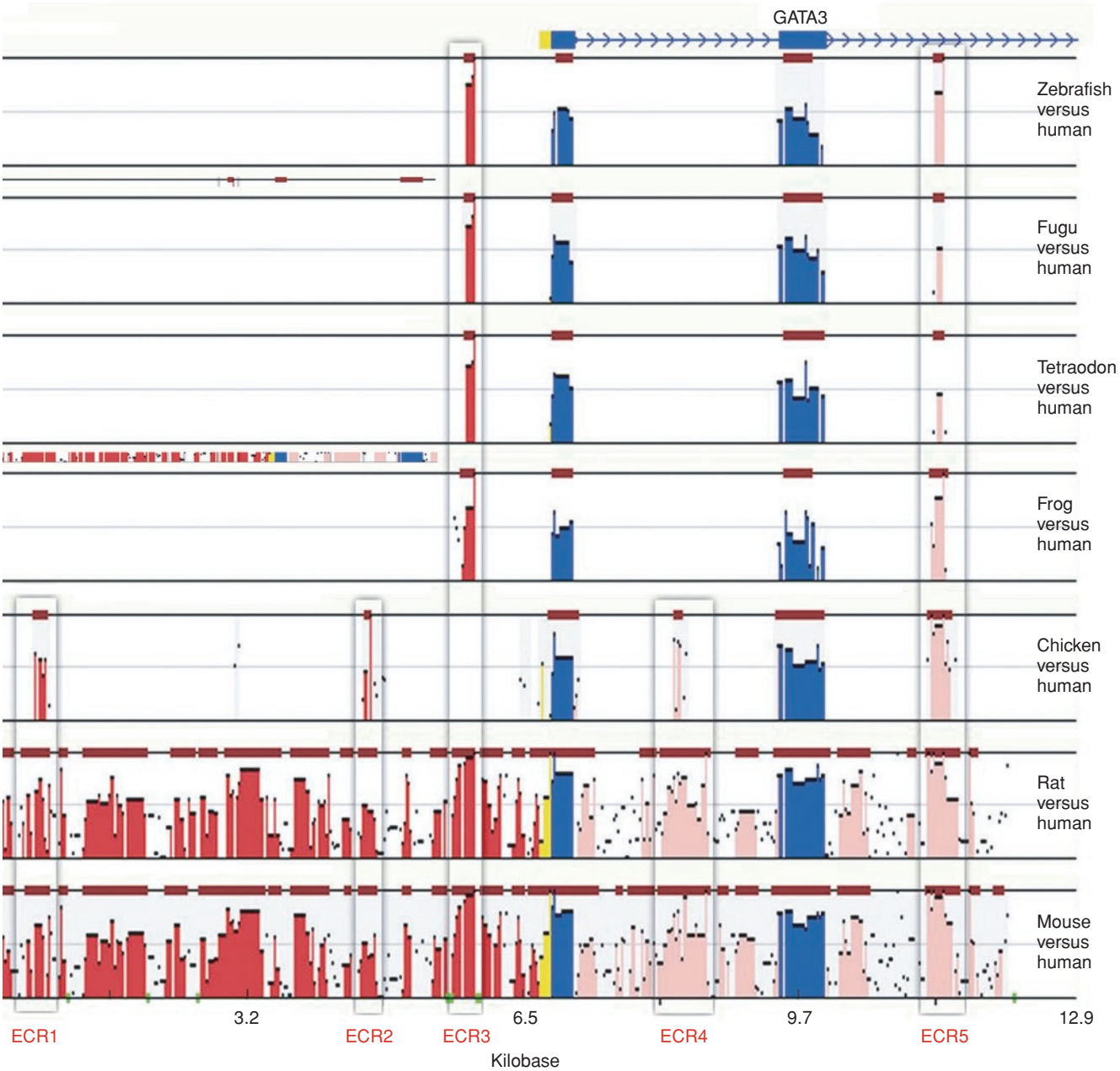
Of the 35 computer scientists who work in Livermore's Biosciences Directorate, one group works alongside biologists, a second works in biodefense research led by the Laboratory's Nonproliferation, Arms Control, and International Security Directorate, and a third is located at the Department of Energy's Joint Genome Institute in Walnut Creek, California.




The Mulan program can generate a phylogenetic tree that depicts the evolutionary relationships between species based on the similarities of their genomes. The red circles represent common ancestors, and the numbers represent nucleotide substitutions per 1 kilobase of genomic sequence.

lack of hair in these species," says Ovcharenko. More interesting, however, was the team's discovery that one RE, called ECR3, is present in all species' genomes except chicken. The team hypothesized that the deletion of this RE many millions of years ago caused the GATA3 gene to function differently in birds, resulting in the hollow bones that make flight possible.

ECR Browser, the most popular of the Livermore online tools, expands the boundaries of alignment to genome scale for comparing vertebrates, invertebrates, and microbes. While offering all the




The Mulan program can align and analyze DNA sequences from multiple species simultaneously. Here, Mulan was used to analyze the GATA3 gene in six species. The program identified five candidate regulatory elements (REs) for the GATA3 gene, which are shadowed in the plot. One RE, called ECR3, is present in all the species' genomes except chicken. It is possible that the deletion of this RE many millions of years ago caused the GATA3 gene to function differently in birds, resulting in the hollow bones that make flight possible.

capabilities of zPicture and Mulan, it also can be used to identify conserved single nucleotide polymorphism (SNP), which is a one-nucleotide genetic change that has appeared in recent evolutionary time (last few thousands of years) and has been found to result in disease or susceptibility to a disease. This ability to identify conserved SNPs is important to medical researchers attempting to identify causes of genetic diseases. Ovcharenko has used ECR Browser to analyze the genomes of human, dog, mouse, rat, chicken, frog, three fish, six fruit flies, and six bacterial species, including *Escherichia coli* and *Yersinia pestis*, the organism that causes plague. ECR Browser is particularly useful for comparing the genomes of bacteria, of which more than 100 have been completely sequenced.

## Comparing Apples and Apples

Comparative genomics works best when comparing "apples to oranges," that is, critical genes and REs that are conserved in genomes of distantly related species such as humans and mice. However, many functional elements in the human genome are not present in the mouse genome. "Comparative genomics fails when searching for functional elements in closely related species or comparing apples and apples," says Loots. "Everything seems to be perfectly well conserved, including elements that are functionally important and those that are not. How can we identify functional elements in comparisons of humans and chimps, whose genomes are more than 98 percent identical? From a practical viewpoint, how can we study primate-specific diseases using comparative genomics?"

Phylogenetic shadowing, a method proposed by Eddy Rubin from Lawrence Berkeley and redefined and implemented by the Livermore team in the eShadow program, provides a better tool for studying closely related species. Phylogenetic shadowing is a statistical method that detects and identifies recently or rapidly evolving ECRs. The program is especially useful for understanding the origin of human-specific genetic diseases.

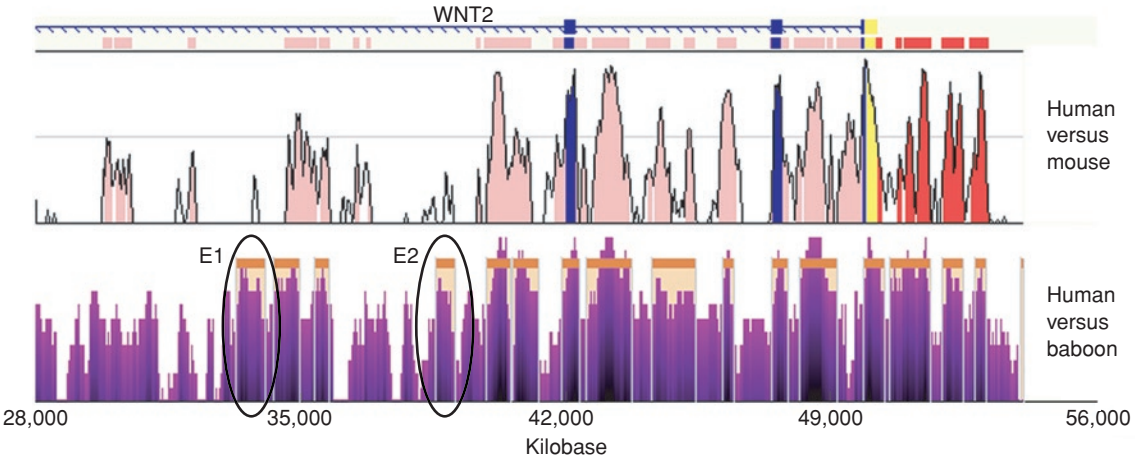
The Livermore researchers used eShadow to study WNT2, an important human gene known to be involved in the early stages of cancer and associated with autism and embryonic development. They found two REs that are highly conserved in humans and baboons but that have diverged in mice. The researchers confirmed the REs' function in the laboratory and found that both reduce the activity of the WNT2 gene. This discovery suggests that the WNT2 REs are conserved in all species, but that different species might use additional elements to modulate the activity of this gene in some organs.

## Characterizing Regulatory Elements

The Livermore comparative genomics toolkit also contains programs to characterize important features of REs and to predict their regulatory activities. Two programs, rVista and multiTF, work with zPicture and Mulan, respectively, to detect evolutionary conserved binding sites for proteins called transcription factors (TFs), which interact with REs. About 10 percent of genes in all species encode TF proteins, yielding at least 2,000 different TF proteins in mammals and about 300 TF proteins in bacteria. TFs combine together and bind to REs at transcription factor binding sites (TFBSs). When a TF complex binds to an RE, the activity of this RE is initiated, and the corresponding gene is switched "on" or "off." Because of the sheer numbers of TFs and their combinatorial action, the variety of protein complexes that can be deployed for gene regulation in any species is enormous. Identifying TFBSs on specific REs is a major challenge and one that is attracting much attention from the bioinformatics community. rVista and multiTF are among the most widely used tools for locating TFBSs.

SynoR, the tool most recently developed by the Livermore team, searches for synonymous gene regulators or groups of REs that share similar gene regulation duties, which often bind to similar TF proteins. When a user inputs a combination of TFBSs, the program scans a genome to find the related REs. The team believes this new tool will be especially useful for establishing links between human diseases and REs. "SynoR has allowed us to make observations that suggest understanding regulation of a particular group of genes could be used to help patients with various immunodeficiencies ranging from inherited syndromes to AIDS," says Ovcharenko.

The interconnection of the Livermore programs with other DNA sequence analysis tools creates a unique portal for studying genomes. For example, some tools are interconnected with the Genome Alignment and Annotation (GALA) database at Pennsylvania State University. Once a region of interest has been found in GALA, a user may examine it using the Mulan tool. Likewise, users can access GALA to find additional information about ECRs found by Mulan. Some Livermore tools also connect to the University of




The eShadow program is particularly useful for determining differences among closely related species. This comparison of the human WNT2 gene with mouse and baboon shows E1 and E2 regulatory elements (circled) that are specific to human and baboon.

California at Santa Cruz's Genome Browser, which includes the completed, annotated genomes of many species and links to other genomic resources.

## Refinements Continue

Ovcharenko is refining the tools to improve their power and utility. Because of the programs' growing popularity, he is planning to adapt the programs, in particular ECR Browser, so they will work on supercomputers.

More scientists worldwide are beginning to research gene deserts and REs. The potential payoff is a more complete understanding of how a cell functions and species have evolved and greater insight into diseases and their possible cures.

*—Arnie Heller*

**Key Words:** Comparative Genomics Center, DNA, ECR Browser, eShadow, evolutionary conserved region (ECR), gene desert, Human Genome Project, Joint Genome Institute (JGI), Mulan, multiTF, regulatory element (RE), rVista, SynoR, transcription factor binding site (TFBS), zPicture.

***For further information contact Ivan Ovcharenko (925) 422-5035 (ovcharenko1@llnl.gov).***